DECISION PROBLEM TO OBTAIN THE OPTIMAL NUMBER OF CLUSTERS IN HIERARCHICAL CLUSTER ANALYSIS

階層クラスター分析における最適クラスター数の決定問題

＜論文概要書＞

A dissertation submitted for the degree of

Doctor of Philosophy at Waseda University

Kimiaki Shinkai

July 2009

一般に, 人間関係や社会構造などの２値化できない情報は, ファジィ理論を適用する事で効果的に分析することができる. ファジィ理論の歴史は, 1965 年, *Journal of Information and Control* に掲載された **L. Zadeh** 教授の論文*"Fuzzy Sets"*から始まり, ここでファジィ集合の概念が提案された.

それ以降, 多くの研究者がファジィ集合論, ファジィグラフ理論, ファジィ推論, ファジィクラスター分析理論をはじめとしたファジィ理論の研究に従事した.

1970 年代には, **E. Mamdani, A. Kaufmann, C. Negoita, T. Nishida**, 1980 年代には, **W. Pedrycz, J. Bezdeck, M. Mizumoto** などにより基礎理論とその応用研究が進められた.

特に階層クラスター分析理論には, クラスタリング手法の分析や, クラスタリング結果の評価分析 (特に最適クラスター数の決定問題) などの興味深い研究がある. クラスタリングの手法の分析に関しては, ファジィ理論では **J. Bezdek, C. Romesburg, M. Anderberg, S. Miyamoto** らが, 多変量解析, 統計解析では **G. Lance, W. Williams, A. Dempster** らが, ニューラルネットワーク理論では **T. Kohonen** らが研究を進めてきた. クラスタリング結果の評価分析 (特に最適クラスター数の決定問題) に関しては, 筆者が AIC 法, ファジィ決定法を考案し応用研究を行った.

実際の階層クラスター分析においては, 多くの場合クラスター数を決定する必要がある. つまり, 分割樹形図 (Fig. 1) においてどのレベルが最適なカットレベルであるかを決める問題である.

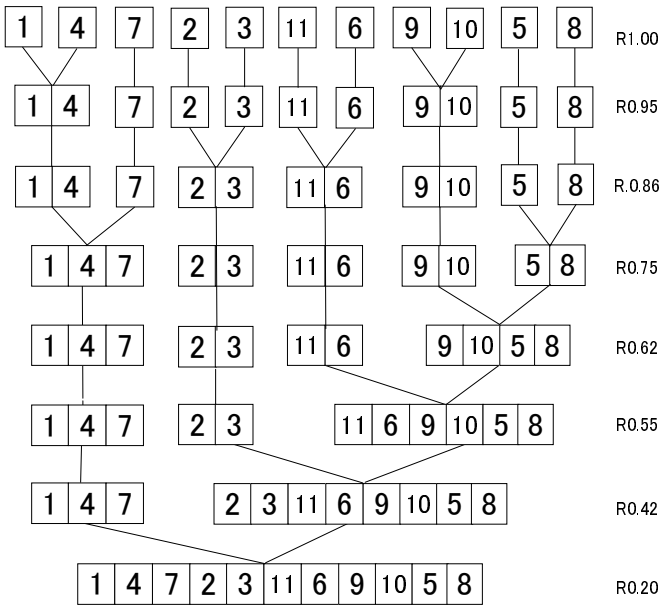

Figure 1: 分割樹形図

この問題に関しては, 従来最急降下勾配法が用いられてきたが, 局所解に陥る可能性があるという問題があった. そこで, 筆者は統計解析における AIC (Akaike's Information Criterion) を応用した方法を考案したが, サンプル数が多くないと最適解を出せないという問題が残った. そこで, サンプル数が少ない場合でも最適解を出せるファジィ決定を応用した方法を考案した.

さらに、この手法をソシオメトリー分析へ応用することを行った. ソシオメトリーは, **J.Moreno** によって提唱された集団構成員の関連構造の測定法で, 構成員相互の友好関係の程度から解析されるが, その構造を明確に表すことは困難であった. ファジィ理論を応用する事で, 集団構成員の構造を明確にあらわすことが容易になった.

本論文では, 最適クラスター数の決定理論を議論するだけでなくソシオメトリー分析を通して, 提案した決定理論の有効性を明らかにする. 実際の章立ては以下の通りである;

第 1 章. 階層クラスター分析手法と性質
第 2 章. 階層クラスター分析の結果の評価
第 3 章. 階層クラスター分析における最適クラスター数の決定理論
第 4 章. ソシオメトリー分析への応用

第 1 章では, 多変量解析に基礎をおくクラスター分析手法について述べる. 階層クラスター分析では, 元データ D を推移律をみたすデータ $\hat{H}$（これと分割樹形図は等価）に変換する. ファジィ理論では通常 **L. Zadeh** の提案した推移包を使う. しかしながら推移包を考えると, $\hat{H}$ の要素の値が D の要素の値からかなり変化してしまうという性質をもつ. この性質を補完するために, 多変量解析において様々な手法が提案されている. そこで, 多変量解析における典型的な階層クラスター分析手法とそのアルゴリズムを紹介するとともに, これらの手法の性質を明らかにする.

第 2 章では, 第 1 章で述べた, 各手法により作られた $\hat{H}$ がどれほど良く D を表しているかの指標について述べる. ミンコフスキー距離は, $\hat{H}$ と D の類似度の度合いを測定する. この値が小さいほど, $\hat{H}$ の要素と D の要素の類似度が高いことを表している. 共表形相関係数は, $\hat{H}$ と D の相関の度合いを測定する. この値が大きいほど, $\hat{H}$ の要素と D の要素の相関が強いことを表している. 実際の階層クラスター分析により得られる種々の $\hat{H}$ の中で, 上記の指標から総合的に判断して, 最適な $\hat{H}$ を得ることができる.

第 3 章では, 階層クラスター分析における最適クラスター数の決定理論について述べる. 実際の階層クラスター分析においては, 多くの場合クラスター数を決定する必要がある. つまり, 分割樹形図 P

においてどのレベルが最適なカットレベルであるかを決める問題である.

この問題に関しては, 従来最急降下勾配法が用いられてきたが, 局所解におちいる可能性があるという問題があった. そこで筆者は, 統計解析における AIC (Akaike's Information Criterion) を応用し方法を考案したが, AIC は最尤推定における漸近正規性を仮定するので, サンプル数が少ないと漸近正規性に関する誤差が大きく, AIC の精度が低くなってしまう. AIC 法は統計分布に基礎をおく道理にかなった方法ではあるが, サンプル数が多くないと最適解を出せないという問題が残った. そこで筆者は, サンプル数が少ない場合でも最適解が求まるファジィ決定を応用した手法 (Fig. 2) を考案した.

分割樹形図の各レベルのクラスターの集合 R_z において, クラスターの分岐度を評価する関数 $p(z)$, クラスターのサイズを評価する関数 $q(z)$ が考えられる. 適度に分岐し, 適度にクラスターサイズの大きい集合 R_{z^*} が全体の状況を良く表していると考え, これをファジィ決定の最大化決定により求める.

また, $q(z)$ に関しては, クラスターサイズとして **Power Mean** という平均型演算を定義し用いている.

$$z* = \max\{z|p(z) \wedge q(z) = \max_{0 \le l \le 1}(p(l) \wedge q(l))\}$$

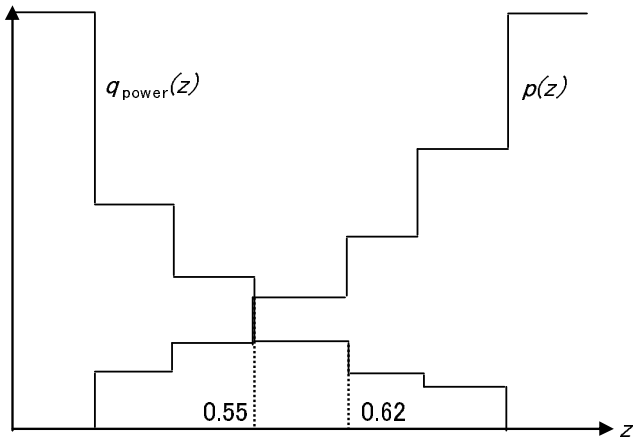

Figure 2: ファジィ決定

さらに評価者が特定したクラスター数やクラスターサイズがある場合, その条件をとりいれた条件付きファジィ決定法も考案した.

第４章では, サンプル数の少ないソシオメトリー分析例を通して, 第３章で論述したファジィ決定法の有効性を明らかにする. 具体的には, ファジィグラフを適用し, （1）最適クラスター数の決定理論 について議論し, その適用例として（2）ファジィソシオグラム分析に関する実践的な適用例 について述べる.

本論文の適用例は, 初等教育における児童の友好関係について考察しているが, 一般に社会集団における人間関係の分析にも広く応用できる.

筆者の提案した階層クラスター分析における最適クラスター数の決定法は, ソシオメトリー分析だけではなく, 教材構造分析や他の様々な意思決定問題にも効果的に適用できる.